Lead Scoring Case Study

Group members :

- Rohan Mahajan
- Roopa Dasari
- Rujit Ravindran

Outline

- Problem Statement
- Solution Approach
  - Data Understanding
  - Data Cleaning
  - Exploratory Data Analysis
  - Data Preparation
  - Model Evaluation - Train Dataset
  - Model Evaluation - Test Dataset
- Conclusion

Problem Statement

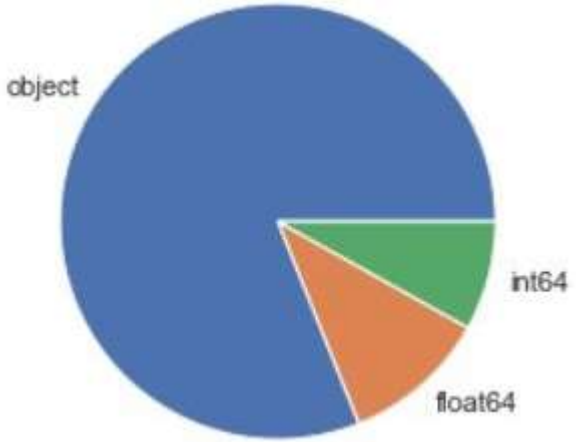
Problem Statement

An education company named X Education sells online courses to industry professionals.The company markets its courses on several websites and search engines.Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead.The company also gets leads through past referrals. Employees from the sales team start making calls, writing emails, etc to leads. By this, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Solution Approach

Data Understanding

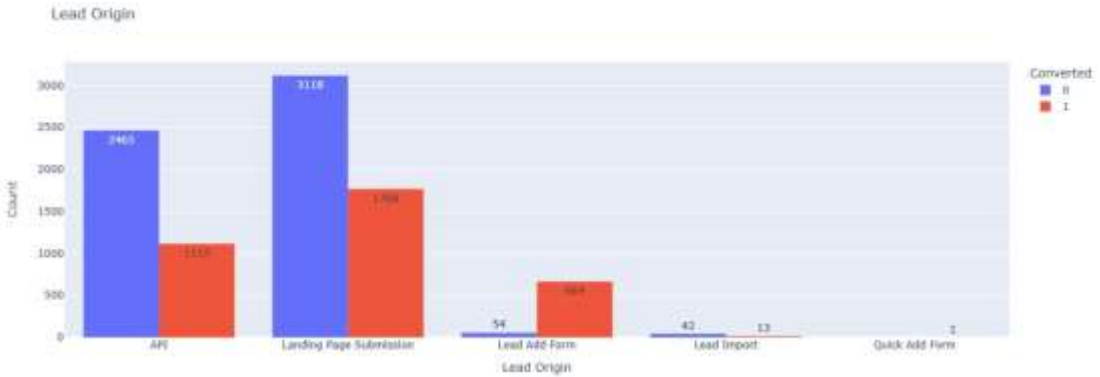


- There are 9240 rows and 37 columns
- There are 17 features have missing values with 5 features having more than 40% missing values.
- There is no duplicate value in this data.

Data Cleaning

- Some features that do not have enough variance have been dropped
- We have also dropped features having more than 40% missing values
- There are some features having one unique value like “Get updates on DM Content”, “I agree to pay the amount through cheque”, “Receive More Updates About Our Courses”, “Magazine”, “Update me on Supply Chain Content” has been dropped.

Exploratory Data Analysis - Lead Origin



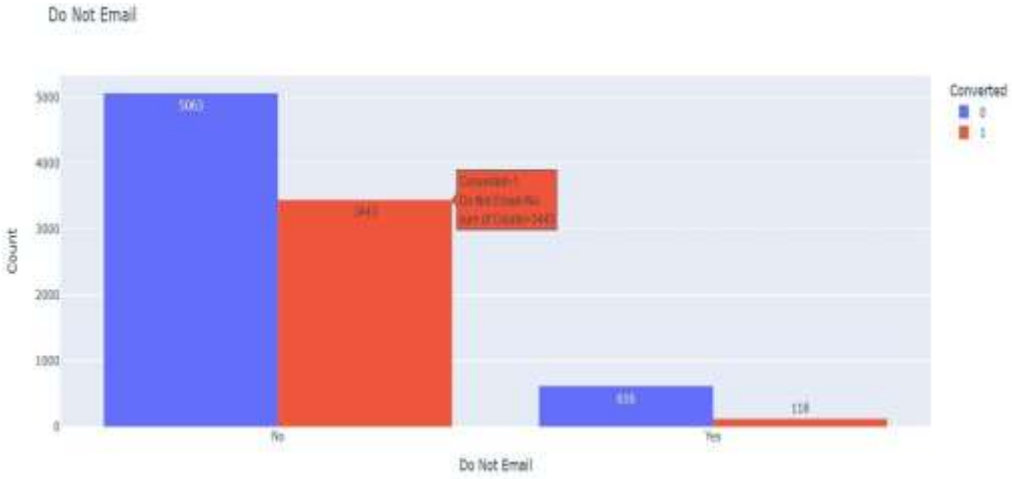
To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

Most Leads originated from submissions on the landing page and around 38% of those are converted followed by API, where around 30% are converted.

Even though Lead Origins from Quick Add Form are 100% Converted, there was just 1 lead from that category. Leads from the Lead add form are the next highest conversions in this category at around 90% of 718 leads.

Lead Import are very less in count and conversion rate is also the lowest

Exploratory Data Analysis - Do Not Email

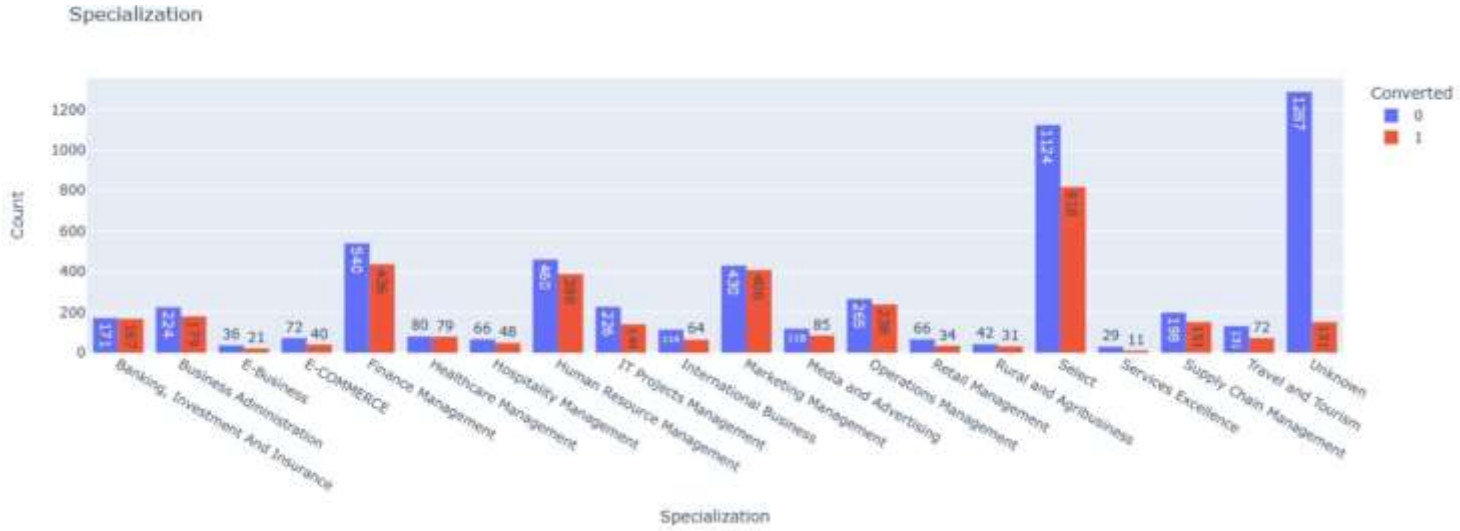


Majority of the people are ok with receiving email (~92%)

People who are ok with email has conversion rate of 40%

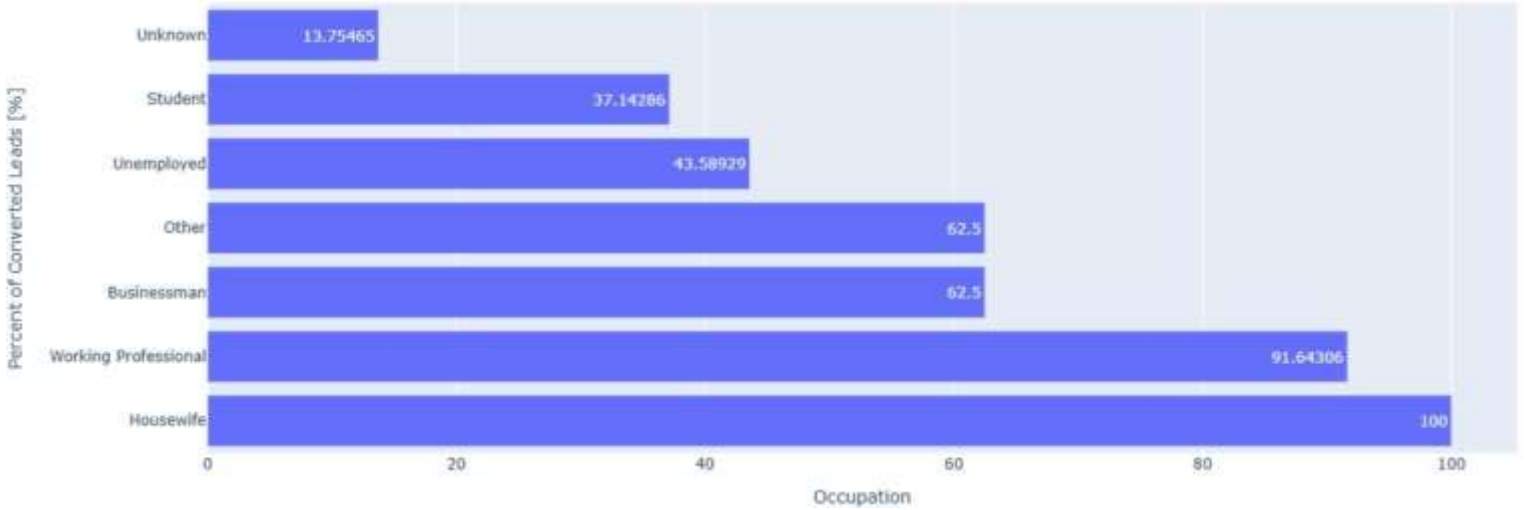
People who have opted out of receive email has lower rate of conversion (only 15%)

Exploratory Data Analysis - Specialization



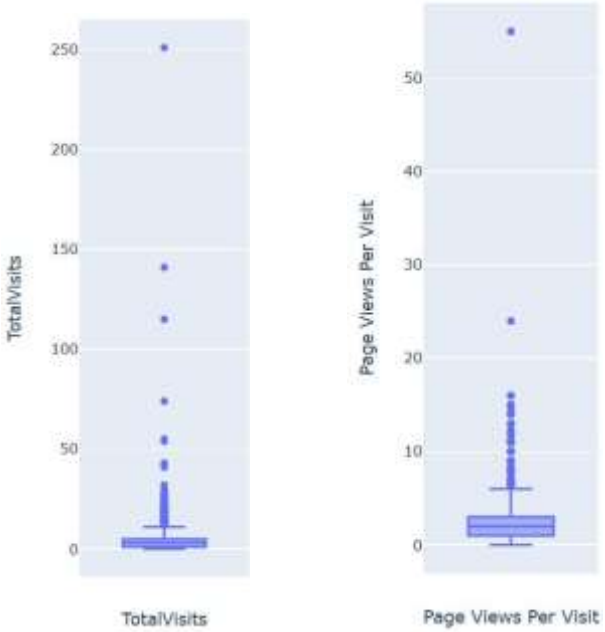
- Most of the leads have not mentioned a specialization and around 28% of those converted
- Leads with Finance management and Marketing Management - Over 45% Converted

Exploratory Data Analysis - Occupation



- Though Housewives are less in numbers, they have 100% conversion rate
- Working professionals, Businessmen and Other category have high conversion rate
- Though Unemployed people have been contacted in the highest number, the conversion rate is low (~40%)

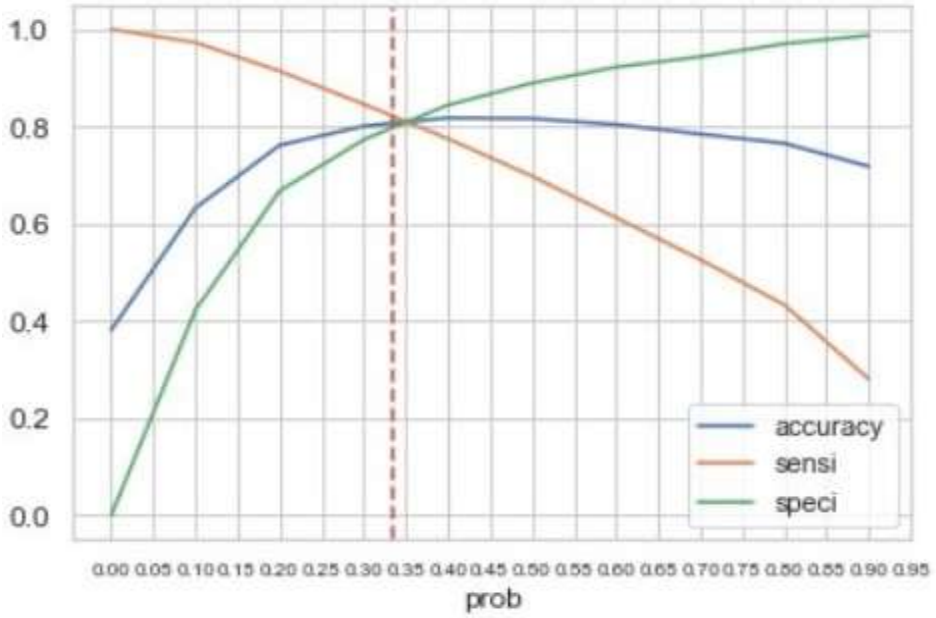
Data Preparation



Though outliers in TotalVisits and Page Views Per Visit shows valid values, this will misclassify the outcomes and consequently create problems when making inferences with the wrong model. Logistic Regression is heavily influenced by outliers. So let's cap the TotalVisits and Page Views Per Visit to their 95th percentile due to following reasons:

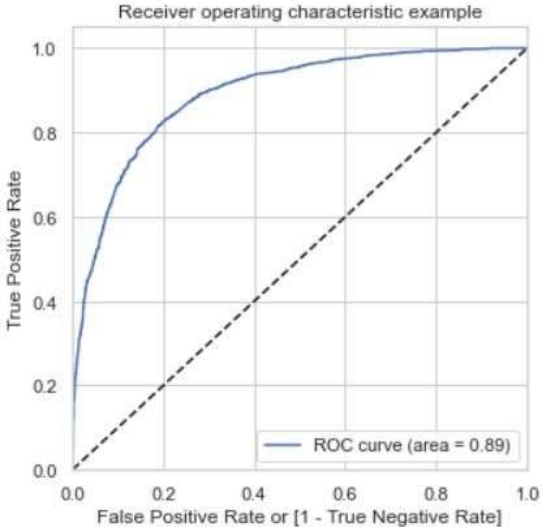
- Data set is fairly high number
- 95th percentile and 99th percentile of these columns are very close and hence impact of capping to 95th or 99th percentile will be the same

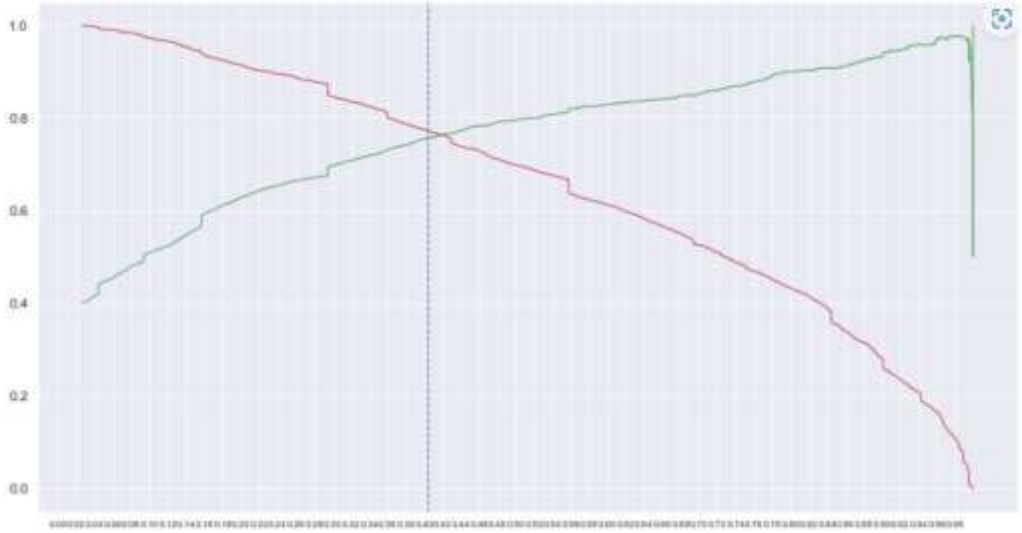
Model Evaluation - Train Dataset



From the plot we can see that 0.335 is the ideal cut-off point.

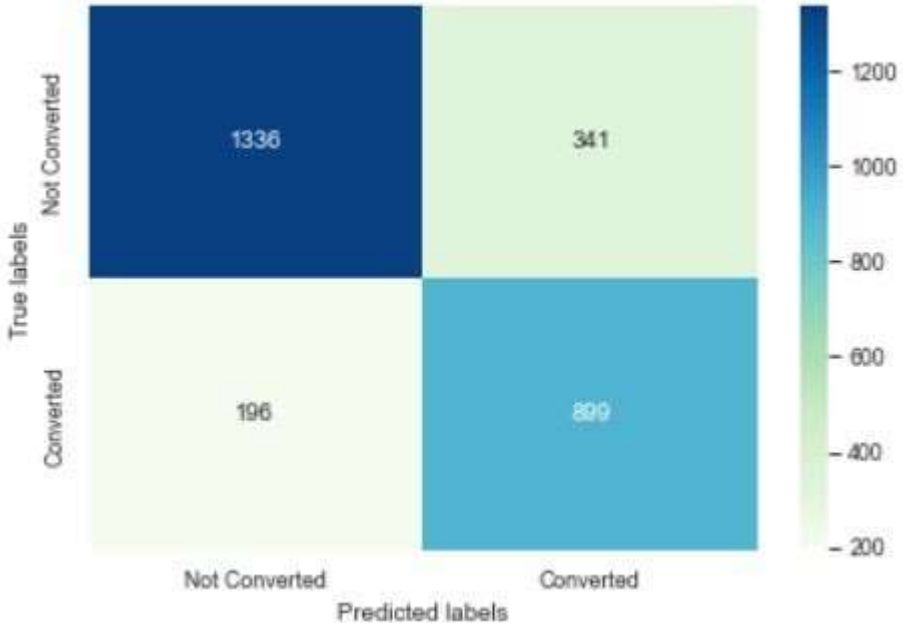
Model Evaluation - Train Data set





- ROC Curve aread is 0.89, which indicates that the model is good.
- Based on Precision- Recall Trade off curve, the cutoff point seems to 0.404. We will use this threshold value for Test Data Evaluation

Model Evaluation - Test Dataset




```
Model Accuracy value is                 :   80.63 %
Model Sensitivity value is              :   82.1 %
Model Specificity value is              :   79.67 %
Model Precision value is                :   72.5 %
Model Recall value is                   :   82.1 %
Model True Positive Rate (TPR)          :   82.1 %
Model False Positive Rate (FPR)         :   20.33 %
Model Poitive Prediction Value is       :   72.5 %
Model Negative Prediction value is      :   87.21 %
```

Conclusion

Conclusion

- Major indicators that a lead will get converted to a hot lead:
- Lead Origin_Lead Add Form : A lead sourced from Lead Origin_Lead Add Form is more likely to get converted
- Occupation_Working Professional :- Working professionals are more likely to get converted.
- Lead_Source_Welingak website : A lead sourced from Welingak Website is more likely to get converted.
- Last Activity_SMS Sent : A lead having SMS sent previously are more likely to get converted.
- Lead Source_Olark Chat : A lead sourced from Olark Chat is more likely to get converted
- Major indicators that a lead will NOT get converted to a hot lead:
- Last_Activity_Olark chat conversation : Customer who had olark chat conversion, are less likely to get converted into hot leads.
- Lead Ongin_Landmg Page Submission : Customer who hadLead Ongin_Landmg Page Submission, are less likely to get converted into hot leads .
- Do Not Email :Customer who choose Do Not Email, are less likely to get converted into hot leads .

Recommendations:

- The company should use a leads score threshold of 34 to identify "Hot Leads" as at this threshold, Sensitivity Score of the model is around 81% which is as good as CEO's target of 80%.